数据流集成分类算法综述 *

许冠英，韩　萌，王少峰，贾　涛

(北方民族大学 计算机科学与工程学院，银川 750021)

摘　要：当前，数据流分类算法的潮流是集成分类算法，因为集成算法提供了比单分类算法更好的性能和更突出的表现。同时在现实世界的实际应用中容易部署，对概念漂移有快速的适应性和恢复性，而且在类不平衡问题的处理中也具有最佳的分类性能。详细介绍了国内外集成分类算法，对集成分类算法的两个部分（基分类器组合和动态更新集成模型）进行了详细综述，明确区分不同集成算法的优缺点，对比算法和实验数据集。并且提出进一步的研究方向和考虑的解决办法。

关键词：数据流分类；集成学习；概念漂移

中图分类号：TP301.6　　doi: 10.19734/j.issn.1001-3695.2018.09.0510

Summarization of data stream ensemble classification algorithm

Xu Guanying, Han Meng, Wang Shaofeng, Jia Tao

(*School of Computer Science& Engineering*, *North University for Nationalities*, *Yinchuan* 750021, *China*)

Abstract: Currently, the trend of data stream classification algorithms is to ensemble classification algorithms. Because the ensemble algorithm provides better performance and more outstanding performance than the single classification algorithm. At the same time, it is easy to deploy in practical applications in the real world, has rapid adaptability and recovery to concept drift, and has the best classification performance in the processing of class imbalance problems. Based on the outstanding features and performance of the above ensemble classification algorithm, it has won extensive research by scholars at home and abroad. This paper introduces the ensemble classification algorithm at home and abroad in detail. The two parts of the ensemble classification algorithm (base classifier combination and dynamic update ensemble model) are reviewed in detail, and the advantages and disadvantages of different integration algorithms, comparison algorithm and experimental data set are clearly distinguished. The paper proposed further research Directions and considerations.

Key words: data stream classification; ensemble learning; concept drift

0　引言

近年来，随着大数据的快速发展，这些数据中蕴含着大量有用的信息，为了获得这些信息，研究人员开展了大量的数据挖掘任务 。最近，在数据流挖掘的研究领域中，从大量快速生成的数据中获得有用的模型已经取得了很多进展。数据流对学习算法提出了若干挑战[2]。学习者的集合已被广泛研究和部署在现实世界的问题中。研究学者们提供了三个理由来证明使用集合而不是单个学习者，即统计学，计算学和代表性[3]。对这种偏好的另一种解释是难以获得强大的学习者，而一组弱学习者相对容易发展并且可以有效地被提升为强大的学习者[4]，只要它们受到了战略训练和结合。集成学习者在数据流设置中很受欢迎，因为除了利用弱学习者之外，它们还可用于处理一般的机器学习问题以及特定数据流的挑战，例如，集合学习者已被广泛应用在解决数据流概念漂移[5]，反复出现的概念[6]，新颖的类检测[7]的问题上。集成学习者在这些问题上都体现出了比单分类模型更好的性能。

和传统的静态数据相比,数据流具有实时的，高效的，快速到达和到达的实例只能处理一次的特点。因此在对数据流中的数据进行挖掘任务时面临以下挑战：a)数据流中的数据仅能处理一次，流动的数据并不能存储在数据仓库当中[8]；b)处理的结果只能最大程度的近似；c)在流中数据的分布会随着时间的推移而改变[9]，即发生概念漂移(concept drift)现象。因此要求面向流处理的算法必须具有快速的恢复性，适应性，准确性和鲁棒

收稿日期：2018-09-11；修回日期：2018-10-26　　基金项目：国家自然科学基金资助项目（61563001）；宁夏自然科学基金资助项目（NZ17115）；北方民族大学研究生创新项目（YCX18055）

作者简介：许冠英（1994-），男，辽宁葫芦岛人，硕士研究生，主要研究方向为数据流集成分类器；韩萌（1982-），女（通信作者），副教授，博士，硕导，主要研究方向为数据挖掘（2003051@nun.edu.cn）；王少峰（1993-），男，陕西人，硕士研究生，主要研究方向为高效用模式挖掘；贾涛（1993-），男，陕西人，硕士研究生，主要研究方向为数据流单分类器.

性。能够实时更新算法，满足算法能处理接下来流中分布改变的数据。在面向处理流数据的算法中,分类是挖掘数据流中最重要也是最关键的部分。目前静态数据处理的方式已经较为成熟,传统分类方法已经不能满足流挖掘任务。对传统挖掘算法来讲，在发生概念漂移的数据流中已经不能进行挖掘任务了，因此面向流数据的处理算法就显得尤为重要。

1 背景知识

1.1 数据流分类

分类(classification)[10]在流数据挖掘任务中是尤为重要的，而且在实际生活中也有很广泛的应用。例如网络入侵检测，金融欺骗，垃圾邮件过滤等问题上[11]。分类任务就是在包含实例和实例所属的类标签中的初始训练集里,通过对数据集中的实例进行学习得到一个目标函数 f, 用这个函数 f 来预测下一个未知实例的类标。即通过某种学习算法在假设样本空间中找到一个 f 的近似函数 g，这个近似函数 g 就叫分类器[12]，也称为分类模型(classification model)。流数据分类任务的输入是记录,每条记录也称作实例或者样本，用元组 (x,y) 表示其中 x 是属性的集合， y 是实例所属的类标签,即样本的类标号。

1.2 增量学习

增量学习（incremental learning）是指一个学习体系不断的从新的样本数据中学习新的知识。在进行流数据分类任务中，需要保证分类器能时刻适应当前流中的数据分布，因此需要获得新数据对原始分类器进行修改，这种不断在线学习新实例的技术（即增量学习）是解决数据流问题不可缺少的[13]。

增量学习主要有两种学习方式，第一种是对原本并不具有增量处理能力的现有算法进行改进，让其具备一定处理数据流中新到来实例的能力。对原始算法进行改造时，核心思想是利用算法的原理或者实验的辅助信息，通过重新进行数学建模从而使算法达到具有增量处理数据的能力。例如有基于支持向量机改造的增量支持向量机（ISVM）[14]和 LASVM[15]。基于随机森林算法改造的在线随机森林（ORF）[16],基于静态广义学习向量量化（GLVQ）的增量学习向量量化（ILVQ）[17]。

第二种方式就是集成增量技术。将集成学习和增量学习相结合，让算法具有增量学习的能力。增量学习对数据流中的实例是非常合适的并且是完美的，只要是数据流中的实例逐个到达，并且学习算法能够从新的数据中学习，同时还能确保之前学习的知识。因此增量学习技术普遍应用在数据流的算法中。例如 Learn++.NC 和 Learn++.UDNC,演化的神经网络 ENN 等。

1.3 分类器模型

随着数据量越来越多，数据挖掘研究人员把分类器模型主要分成两类。单分类器模型和集成分类模型。周志华教授在机器学习书中提出了集成学习是目前机器学习最具有前景的机器学习技术之一，而且集成学习模型对单分类模型的优势突出，对概念漂移的发生体现了快速的恢复性，适应性，并在分类准确度上，比单分类器精度更高。

1.3.1 单分类模型

单分类器模型是不断的用新到来的数据来递归的更新自身结构，使自身结构能够适应流中数据的变化，并能对流中的实例准确分类。单分类模型主要基本技术有 KNN,决策树,SVM,贝叶斯，逻辑回归和神经网络等。

a)KNN。找到训练集样本空间中的 K 个距离预测样本 x 最近的点,统计 K 个距离最近的点的类别,找出个数最多的类别,将 x 归入该类别。优点：思想简单，理论成熟，既可以用来做分类也可以用来做回归；可用于非线性分类； 训练时间复杂度为 O(n);缺点：计算量大；难以处理类不平衡问题（即有些类别的样本数量很多，而其他样本的数量很少）。

b)决策树方法采用自顶向下的递归方式，在决策树的内部节点进行属性值的比较并依据不同的属性值推断该节点向下的分支，在决策树的叶节点得到结论（预测）。决策树是一个类似于流程图状的树结构，决策树中每一个内部节点表示在一个属性上的测试，每一个分支代表一个测试输出，而每一个叶节点代表类或类分布。优点：计算量简单，可解释性强，比较适合处理有缺失属性值的样本，能够处理不相关的特征。缺点：容易过拟合。

c)SVM 法即支持向量机(Support Vector Machine)法，是二元分类模型。SVM 的主要思想可以概括为两点:（a）它是针对线性可分情况进行分析，对于线性不可分的情况，通过使用非线性映射算法将低维输入空间线性不可分的样本转换为高维特征空间使其线性可分，从而使得高维特征空间采用线性算法对样本的非线性特征进行线性分析成为可能;（b）它是基于结构风险最小化理论，在特征空间中构建最优超平面，使学习器得到全局最优解，并且在整个样本空间的期望以某个概率满足一定上界。优点：可用于线性/非线性分类，也可以用于回归；低泛化误差；容易解释；计算复杂度较低。缺点：对参数和核函数的选择比较敏感。

d)贝叶斯分类器的分类是通过对象的先验概率，利用贝叶斯公式计算出其后验概率，所谓的后验概率也就是该对象属于某一类的概率，然后选择具有最大后验概率的类作为该对象所属的类。即在哪个类标上的后验概率大就属于哪个类。优点：对小规模的数据表现很好，适合多分类任务，适合增量式训练。缺点：对输入数据的表达形式很敏感（连续数据的处理方式）。

e)逻辑回归（Logistic Regression）是用于处理因变量为分类变量的回归问题，常见的是二分类或二项分布问题，二分类问题的概率与自变量之间的关系图形往往是一个 S 型曲线，采用的 *Sigmoid* 函数实现。优点：实现简单，分类时计算量非常小，速度很快，存储资源低。缺点：容易欠拟合，一般准确度不高，只能处理二分类问题。

f)神经网络就是一组相互连接的输入输出神经单元，这些单元之间的每个连接都关联一个权重。在网络学习阶段，网络通过调整权重来实现输入样本与其相应类别（正确）的对应。由于神经网络学习主要是针对其中的连接权重进行的，因此神

经网络的学习有时也称为连接学习。优点：有很强的非线性拟合能力，可映射任意复杂的非线性关系，而且学习规则简单，便于计算机实现。缺点：不能向用户提出必要的询问，而且当数据不充分的时候，神经网络就无法进行工作。把一切问题的特征都变为数字，把一切推理都变为数值计算，其结果势必是丢失信息。

g)单一模型结构复杂，表现能力差，但是单一模型稳定性很好，可塑性较高。对发生概念漂移的流数据也有很好的表现。例如最早提出的 VFDT[18],它使用满足 hoeffding 边界的少量数据训练出来的决策树和使用不满足 hoeffding 边界的大量数据训练出的决策树能有近似的分类结果。由于当时的局限性，因此没有考虑到概念漂移现象。但是，目前许多集成算法还是保留用 VFDT 作为训练算法。可见 VFDT 单分类性能还是比较突出的。在 VFDT 基础上，提出了 CVFDT[18],目的是为了解决数据流中概念漂移的问题。

1.3.2 集成分类模型

在平稳的数据流情况下，首先将训练数据分成不同的训练子集，在每个子集上，采用某种学习算法对子集上的数据进行学习，每学习一个子集，就在对应的子集生成一个基学习器（基分类器），然后采用某种组合方式，将多个基学习器组合成集成学习器（集成分类器）。集成分类器在预测实例的类标签时，通过某种机制将各个基分类器的结果进行综合，最后将综合以后的结果进行输出，得到未知实例的类标签（预测）。集成学习把多个学习器结合起来，因此本文需要考虑这样下面这个问题，怎样才能让集成学习器体现出比单一学习器更好的学习性能？所以，解决这个问题就出现了如下两种解决思路。要获得好的集成，个体学习器应好而不同，即个体学习器要有一定的准确性，学习器不太坏；还有要有多样性（diversity），即学习器间具有差异性。要想获得准确率的提升,必须要使集成分类器中的基分类器彼此间有一定的相异度。这可以通过让基分类器的训练数据不同，甚至使用不同的基分类器算法来实现。

考虑二元分类问题，$y\in\{-1,+1\}$和真实函数 f，假设基分类器的错误率为 ε，对每个基分类器 hi 有[19]：

$$\mathrm{P}(h_i(x)\neq f(x)=\varepsilon$$

假设通过简单多数投票法结合 T 个分类器，超过半数的基分类器正确，则集成分类正确性：

$$\mathrm{H}(x)=\mathrm{sign}\left(\sum_{i=1}^{T}h_i(x)\right)$$

假设基分类的错误率相互独立，则由霍夫丁不等式可知，集成的错误率 ε：

$$\mathrm{P}(\mathrm{H}(x)\neq f(x))=\sum_{k=0}^{\lfloor\frac{T}{2}\rfloor}(C_T^K)(1-\varepsilon)^K\varepsilon^{T-K}\leq\exp\left(-\frac{1}{2}T(1-2\varepsilon)^2\right)$$

上述不等式明确表明：随着集成模型基分类器个数 T 的增加，集成模型的错误率呈现指数级的下降，最终趋近于 0。集成学习就是把所有个体学习器的结果做简单的投票，这样就能获得比个体学习器更好的泛化性能。这样说需要满足一个关键假设：基学习器的误差相互独立。但是在现实任务中，个体学习器是为解决同一问题训练出来的,它们显然不可能相互独立。事实上，个体学习器的准确性和多样性本身就存在冲突。一般的，准确性很高之后，要增加多样性就要牺牲准确性。

2　集成分类器中的概念漂移问题

2.1　概念漂移的定义

在进行流数据挖掘任务的时，目标概念会随着时间和周围环境的变化而发生巨变。不变的概念也会发生巨变。例如用户浏览网站的倾向会受到实时热点新闻，个人喜好变化的影响而改变。定义这种目标概念随着非确定性因素而发生改变的现象称作概念漂移。一种关于概念经典的定义中，将概念定义成一组对象的集合。但是这个概念的定义并不能用于流数据。目前大多数关于概念漂移的文献都是采用先验概率，条件概率，后验概率来定义概念漂移的。文献[20]分析了概念漂移发生的三种形式：

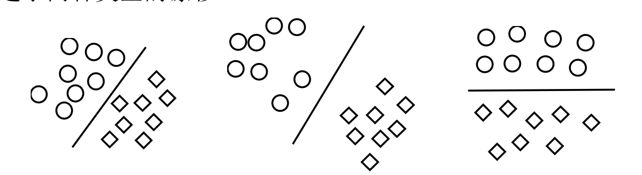
a)类的先验概率 $P(c)$会随着时间的改变而改变。

b)一个类或者几个类的条件概率 $\mathrm{P}(X\mid C_i)$，$i=1,2,3,\ldots,m$ 可能会随着时间的推移而发生改变。

c)后验概率 $\mathrm{P}(C_i\mid X)$，$i=1,2,3,\ldots,m$ 的改变被认为是真正的概念漂移，即同一个实例在不同的时间域中，具有不同的类标签。

2.2　概念漂移的类型

根据类标号的先验概率，条件概率和后验概率。概念漂移的类型主要分为虚假概念漂移和真实概念漂移[21]。前一类概念漂移不影响决策边界（后验概率），但影响条件概率密度函数。因此，它不应该直接影响所使用的分类器。后一类概念漂移对决策边界（或后验概率）有影响，并可能影响条件概率密度函数。这种类型的变化可能会，明显影响分类器的性能。图 1 描述了两种类型的漂移[22]。



(a)初始数据分布　(b)虚假概念漂移　(c)真实概念漂移

(a)Initial data distribution　(b) False concept drift　(c)Real concept drift

图 1　概念漂移两种类型的不同边界

Fig.1　Concept drifts two types of different boundaries

2.3　概念漂移的处理技术

目前处理概念漂移的方式有许多种，包括滑动窗口模型，概念漂移检测器，集成学习模型，在线学习者等[23~25]。

a)滑动窗口。滑动窗口技术主要是保留了一个缓冲区，在缓冲区中最新的实例认为是最能体现当前数据流中数据的分布状态。它们用于训练和更新模型。并且一旦新的实例到达，之前的实例就被丢弃。通过储存最新的状态来更新实时的数据流。

滑动窗口提供了一种只分析数据流中最新数据元组的途径。这种技术只考虑当前窗口内的数据,并不需要对数据随机抽样也不保留过时数据的统计信息[26,27]。代表算法有 ADWIN Bagging[28]和 Leveraging Bagging[29]、ADWIN2、SERDRIFT[30,31]、ECISD[32]。

b)概念漂移检测器。它们可以被视为与给定分类器结合的外部算法。它们的目的是监视数据流的特定属性，例如标准偏差[33]，预测误差[34]或实例分布[35]。假设这些特征的任何变化都是由漂移存在引起的。因此，通过测量变化水平，探测器能够检测报告进入的水平变化。代表算法 DDM[36]、EDDM[37]。

c)在线学习者。按照在线处理实例的方式更新模型，从而在流发生时尽快调整流。这样的学习者必须满足一系列要求[38]：每个对象在训练过程中必须只处理一次，处理每个实例的计算复杂性必须尽可能小，并且其准确性不应低于在批量数据上训练的分类器。代表算法有 CUSUM[39]FIMT-DD[40]。

d)集成学习者。使用组合方式的集成学习者因为具有多样性和复杂的结构，每个单分类器都有良好的性能，它们可以轻松适应流的变化，提供灵活性和预测能力的增益[41]。两种主要方法假设一个变化的集合[42]或更新基分类器[43]。新分类器正在训练最近到达的数据（通常以块的形式收集）并添加到集成模型中。修剪用于控制基本分类器的数量并删除性能最差或最旧的模型。代表算法有 DWM[44]、AWE[45]、SEA[46]、EB[47]、OCBOOST[48]、OAUE[49]

e)其他技术。在文献[50]中，作者利用 Kappa 系数的范围来进行检测，作者认为 Kappa 系数为 65％是可接受的和谐。当每个输入包的最后 100 个样本的 Kappa 系数小于 65％时，所提出的方法将分类过程称为“随机”。发生这种情况时，会替换加权函数并丢弃不良分类器。代表算法文献[50]、ASHT[51]。

3　集成分类算法

集成分类学习是通过集成多个基分类器共同决策的机器学习技术，通过调用简单或者复杂的增量学习算法，获得多个性能好而不同的基分类器，然后采用某种结合方式将全部基分类器组合成一个集成分类器。在 1.3.2 节介绍了集成的理论和发展方向，因此如何产生并结合好而不同的个体学习器，就是集成学习研究的核心，怎么集成？集成什么样的个体学习器？根据个体学习器的生成方式，目前的集成学习方法大致可分为两大类[52]：a）个体学习器间存在强依赖关系、必须串行生成的序列化方法，代表是 Boosting[53]；b）个体学习器间不存在强依赖关系、可同时生成的并行化方法，代表是 Bagging[53]和随机森林（Random Forest）[54]。

对集成学习者来讲，处理不同的问题需要不同的基准算法，虽然问题的本质上都是为了追求良好的分类性能，但是根据分类的具体问题选择合适的基础学习者是获得准确集成分类器的必要前提。分类器通常可以自然地仅处理一种类型的特征域而无须求助于输入的预处理。因此，假设所有特征具有相同的域，则可以根据输入特征域选择基础学习器;使用处理离散和处理连续特征的基础学习器。例如广泛使用的 Hoeffding tree。因为 Hoeffding 边界确定仅仅需要少量的数据就可以训练出和全部训练数据的近似，并且在实验中有着良好的分类性能。用 Hoeffding tree 当做基准算法的有：ASHT、HWT[55]、AWT-ADWIN[56]。在处理数据流的问题中，CVFDT 是可以处理概念漂移的一种快速决策树基准算法，例 CVFDT Update Ensemble(CUE)[57]算法。通常用于集合流学习的其他基础学习者包括朴素贝叶斯，支持向量机，和多层感知机等等。

集成分类模型分成基分类器的组合和模型的动态更新两个部分。

3.1　基分类器的组合

集合成员整体的预测要比单个分类器预测的性能高，本文寻求一种适当的方法来组合这些基分类器。目的是为了能更好的区分那些较难辨别的类。从前学者们致力于研发精度更准确的单分类模型，从没有研究过分类器组合的预测[58]。在组合的方式中，用到的最多的是投票法和固定基分类器的集合。

3.1.1 投票法

投票法是在集成模型预测时，如何选择单个分类器的输出结果的方法。目前主要的投票法分为多数投票法、加权投票法和其他投票法。

a）多数投票法。多数投票法是初始化全部基分类器相同的权重，在进行最终预测时，若某标记的基分类器得票超过半数或者自定义得票最多的基分类器被认定为最后的预测。若同时有多个得票最高的基分类器，那么从它们中随机选取一个。使用多数投票的数据流集成分类算法包括：online Bagging and boosting、MOSOB[59][60]、OOB 和 UOB[60]等。在线算法的主要思想是数据不是基于块状到达的，而是数据流中的实例一个一个单独到达的，学习算法在线处理每个单独实例。当数据量 N 趋近于无穷，将满足泊松为 1 的分布。

b）加权投票法。由于各个基分类器性能的差异，因此根据集分类器的表现对基分类器进行加权。从而着重看待那些分类性能好的基分类器，以便在整体预测中，能输出正确结果。

简单一点的加权就是根据分类器的准确性。例如 AWE(accuracy weighted classification)，AWE 算法是面向数据流的基准算法,AWE 根据基分类器的均方误差，分配反比其均方误差的权值，采用多折交叉的方法计算权值，权值大的替换权值小的基分类器进行模型更新。Weighted majority (WM)和 matrix multiplicative weights(MMW)[61]算法，WM 根据过去的表现对分类器的预测进行加权，这样每个分类器都有一个权重 β，每当不正确地预测时 β 就会减少。Accuracy Update Ensemble(AUE)计算最新数据块上的分类器和集成中全部的基分类器对最新数据块分类的误差，对比性能。若集成模型中某一个基分类器的误差比在最新数据块上的基分类器的误差大，那么替换最差的基分类器。

复杂一点的加权 online accuracy update ensemble (OAUE)算

法和传统的基于块的方式不同，它是基于块和增量的学习算法。在 AUE 加权公式中，引入了时间的概念，把加权公式变为随着时间流动的增量加权表达式。CVFDT update ensemble(CUE)使用 VFDT 作为训练基分类器的算法,在模型更新的过程中，选择 $MES_i > MSE_r$ 的基分类器，作者基于这样的想法：MSE_r 表示任意一个基分类器对最新数据块中的实例随机猜测的误差，如果基分类器的均方误差 MES_i 大于 MSE_r，说明基分类器 C_i 的准确性比随机猜测的准确性还低，这样的基分类器对模型是没有贡献的，需要用最新数据块的数据更新这些基分类器,为了增加基分类器之间的相异度，在最新数据块对基分类器进行更新的过程中，对数据块中的数据采用 bagging 操作，由于基分类器的训练数据很大程度上不同，因此增加了基分类器彼此之间的相异度。

处理类不平衡的加权 ensemble classifiers for imbalanced data stream(ECISD)算法使用 AUE2[62]算法作为基准算法，因为是类不平衡集成分类算法，首先采用 SMOTE 过采样和 Tomek -links 欠采样相结合的方法对原数据进行采样，在基分类器的加权上，引入了代价的概念。模型更新过程中是利用基分类器对模型准确率的贡献来淘汰性能最差的基分类器。

其他加权算法主要还有 adaptive classifiers-ensemble (ACE)[63],weighted ensemble online bagging (WEOB) [64]。

在算法 CUE,AUE,WM,AWE,ECISD,AUE2 都是采用均方误差来分配基分类器得权重。其均方误差表达式为

$$MSE_{ij} = \frac{1}{B_i} \sum_{(x,y)\ni} \left(1 - \int_y^j (x)\right)^2$$

随机分类器得均方误差表达式为

$$MSE_r = \sum_y p(y)(1 - p(y))^2$$

在 CUE,AWE,WM 算法中基分类器权重为

$$W_i = \frac{1}{MSE_i + \varepsilon}$$

在 AUE,AUE2 算法中基分类器权重为

$$W_i = MSE_i - MSE$$

在 ECISD 算法中基分类器权重为

$$W_{ij} = \frac{1}{Ct_{ij} + MSE_{ij} + MSE_r + \alpha}$$

其中 Ct_{ij} 是基分类器 C_j 在数据块 B_i 上误分类的总代价为

$$Ct_{ij} = \frac{1}{|B_i|}(x,y)\ni B_i \sum_{y'} J_y \quad y^{'}(x) \cdot \int_{y'}^{j} (x)$$

在算法 OAUE 中基分类器均方误差表达式加入了时间的概念：

a) $MSE_i^t = MSE_i^{t-1} + \frac{e_i^t}{d} - \frac{e_i^{t-d}}{d} \quad t - \tau_i > d$

b) $MSE_i^t = \frac{t - \tau_{i-1}}{t - \tau_i} MSE_i^{t-1} \quad 1 \le t - \tau_i \le d$

c) $MSE_i^t = 0 \quad t - \tau_i = 0$

其中：

$$e_i^t = \left(1 - \int_{ij}^{t} (x^t)\right)^2$$

$$MSE_r^t = MSE_r^{t-1} - r^{t-1}(y) - r^{t-1}\left(y^{t-d}\right) + r^t(y) + r^t\left(y^{t-d}\right) \quad t > d$$

$$MSE_r^t = \sum_y r^t(y) \qquad t = d$$

$$r^t(y) = p^t(y)\left(1 - p^t(y)\right)^2$$

$$w_t^i = \frac{1}{MSE_r^t + MSE_i^t + \varepsilon}$$

c）其他投票法。考虑到多数投票的缺陷，作者[65]提出了最新的一种投票方式，在投票阶段，并不是全部的基分类器全部参与投票，而是设定一个弃权阈值，阈值为 0.65。即如果基分类器的准确性低于 0.65，那么在决策阶段就不参与投票，准确率高于 0.65 的基分类器参与投票。Dynamic weighted majority(DWM)即动态加权多数算法，DWM 维护可变数量的基分类器。集成模型预测由加权多数票决定。每个分类器的权重在每次正确预测时都会增加，否则会减少。如果分类器的权重低于给定阈值，则从集合中移除分类器。如果集合决策不正确，则添加新专家。由于这种集合更新策略可以为噪声数据流产生过多的添加和删除，因此作者引入了一个参数 β 来确定将在多少个实例中进行整体更新。

Modal mixture model(M3)[66]是基于异构模型类型的加权多数集成算法，其中模型权重使用强化学习技术在线更新。因为它使用基分类器的混合，并通过强化学习借鉴概念的在线方法调整整体成员的权重。当数据流中的数据点进入应用程序时，它们最初用作测试数据以评估实验报告的整体算法。接下来，可以选择每个数据点（通过均匀随机选择）以用作训练数据。训练数据点用于单独训练每个基分类器（VFDT 和朴素贝叶斯模型相混合），并使用简化的训练准确性来更新每个基分类器的权重。

Droplets ensemble algorithm(DEA)[67]算法，这是一种全新的集成学习算法，在 2016 年 ICDM 上获得最佳论文奖。它动态地保持 n 个 BL（base learner）($F = f_1, \ldots, f_n$)的集合以及与当前概念相关的 p 个 $Droplet$ $\left(MAP = \{D_1, \ldots, D_p\}\right)$ 集合。BL 可以是任何基础算法，只要他们能在具有概念漂移的数据流上进行分类。

Droplet 本质就是一个多维特征空间，每一个 *Droplet* 的 D_t 与观测值 x_t 相关联。并保持指向 BL 的指针。首先本文需要找到与 BL 相关联的最新的特征子空间 D_t ,通过对每个 BL 在最近 N 个 *Droplets* 的预测误差相加来完成的。如果唯一的 BL 单独最小化这个和，那么它与最新的 *Droplets* 相关；否则（如果至少 2 个 BL 最小化预测误差的总和），搜索空间依次扩展为 N

$+1; N+2; N+3;$ :::离 *Droplets* 最近，直到找到一个最好 *BL*。然后将新的 *Droplet* D_t 添加到 N_{norm} 坐标处的特征空间中。存储预测误差的矢量 e_t+1 并创建指向上一步中找到的最佳 *BL* 的指针。然后该算法经过重叠的 *Droplet* OD_t 的集合，如果它不是空的，它减少了在 OD_t 中输出错误预测的 *Droplet* 的影响。这是通过缩小它们的半径来完成的，这将使它们不太可能预测在特征空间的该区域中接收的未来观察。如果内存已满，算法使用 3 个不同的标准来选择将被删除的 *Droplet*：

1.移除半径最小的 *Droplet*。2.如果所有 *Droplet* 具有相同的半径，则移除已输出最大错误预测数的 *Droplet*。3.如果标准 1.和 2.失败，请删除最旧的 *Droplet*。

3.1.2 固定集合体系结构

固定集合体系结构定义了基分类器如何相互协调和工作。大体上有三种不同的体系结构，线性和非线性组合（例如加权投票），级联和网络。级联是一种分类器的输出包含了多个分类器的输入的框架（例如 stacking）。网络是一种分层的框架，将成员排列成树状结构的集合或网络的集合。将给定的集合结构分类为：简单的线性组合、元学习器和分层的树状结构。

a)线性和非线性组合。基分类器在输入数据上进行训练，决策融合阶段由组合函数进行投票。主要的算法有 online accuracy update ensemble(OAUE)、online Bagging and boosting、Leveraging Bagging,在文献[68]中提出了基于线性和非线性的加权组合，首先将数据流分成数据块，在每个数据块上训练基分类器，所提方法中的加权过程是在基本分类器上进行的;当在不同条件下添加输入数据时，使用一个线性函数和一个非线性函数。当概念是静止时，非线性函数更有效，而当输入数据的波动值得注意时，线性函数是优选的。另一方面，没有漂移的非线性函数使分类器免受噪声和无关数据的影响。其中权值分配用的是平均绝对误差 *MAE*。

$$MAE=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-y_i\right|$$

线性加权函数： $W_{Linear}=max\{MAE_r-MAE_i \quad ,\varepsilon\}$

非线性加权函数： $W_{Non_linear}=\dfrac{1}{MAE_r+MAE_i+\varepsilon}$

b)元学习器。当训练数据非常庞大时，元学习被称作是 更强大的组合策略。元学习器是通过另一个学习器来进行结合。个体学习器称为初级学习器，用于结合的学习器称为次级或元级学习器。经典的代表就是 Combining Restricted Hoeffding Trees using Stacking[69]，算法是利用列属性子集来构建一组 hoeffding tree，然后利用数据流的 ADWIN 监测机制设置 *sigmoid* 感知器的学习速率，当感知器分类性能不佳时，使用 ADWIN 重置集合成员。

c)分层结构。在该结构中，集合成员被表示为网络的顶点，其连接根据特定标准确定。分类器之间的连接根据无标度网络模型生成，使得具有较高估计准确度的分类器更可能连接到最近添加的分类器。在投票期间，分类器的权重与给定的中心度量 α 成正比，例如，特征向量，中介度等。由于高度准确的基分类器通常需要接收大多数连接，因此预期这些连接对整体决策具有更高的影响。在 social adaptive ensemble（SEA）[70]和 advances on the social adaptive ensemble(SAE2) [71]中，每对学习者根据相似性函数进行连接和加权。由所有这些连接形成的加权网络在每个周期更新，以更好地近似学习者的当前状态。在预测期间使用该网络布置，其中首先在类似分类器的子集内组合各个决策，然后组合这些子集决策以获得最终预测。

为了更方便地分析算法性能和优缺点，在表 1 中详细介绍了算法的数据集、对比算法和优缺点。总体来说 AWE 是面向数据流集成分类器的最经典的算法，它开创了一个时代随后出现的许多算法都是基于 AWE 的加权思想，但是早期的算法性能并不突出，对概念漂移的处理效果不明显。目前性能较好的，且在实践中可以应用到多个场景的算法大致有以下几种：AUE2、DEA、SEA2、WEOB1、WEOB2。

3.2 模型动态更新

集成分类算法另一个重要的部分就是如何对集成模型进行更新，本文当然希望留下能够适应当前数据分布的基分类器，删除性能较差或者较旧的基分类器，因此在进行数据流分类任务时，从数据流中学习需要的算法不仅仅是要求精度，还要有快速适应环境和恢复环境的能力。对概念漂移的适应性和恢复性是对分类器性能的一个重要评价。所以动态更新集成分类器就是重中之重。

3.2.1 增量模型

在 2.2 中介绍了数据流中增量学习的两种方式。因此，在这节主要介绍典型的增量集成算法，并且比较了增量模型和批处理（基于数据块）算法。批学习者在使用数据流中的实例进行训练之前，必须需要储存一批的实例，把数据流分成不同的数据块，然后在每个数据块上进行训练。每当最新的数据块到达时，用这个最新的数据块对集成模型进行更新，通常更新的一般策略是比较候选分类器和集成模型中全部基分类器的性能，利用某种评价得到基分类器和候选分类器的性能差异，淘汰或删除性能最差的基分类器。增量学习者在实例到达时对实例逐个进行训练，通常来讲增量学习者在应用于呈现逐渐或渐进漂移的流或与漂移探测器结合使用时更有效。在突然漂移的情况下，增量学习器（没有漂移探测器的帮助）可能需要更长的时间来恢复，因为这种模型受到先前提出的概念影响，而批量学习器则完全丢弃其先前的模型。由于数据流的特点，所以增量学习算法是数据流分类中必不可少的一部分。Learn++算法是一种非常典型的增量集成算法，它是采用加权投票的方式进行最终的预测。在 Learn++算法的基础上，根据解决实际问题的不同需求出现了以下几种算法：Learn++.MT，Learn++.MT2，Learn++.NSE，Learn++.SMOTE 等。增量学习者的例子还包括贝叶斯分类器，决策树，回归树等 [72~75]。

表 1 算法性能比较

Table 1 Algorithm performance comparison

算法	实验数据集	对比算法	优缺点
MOSOB	Gearbox、Smart Building、PAKDD iNemo、KDD	OOB、UOB、RLSACP、WOS-ELM	缺点：MOSOB 中的搜索技术是一种蛮力方法，在一组有限的候选者中寻找最优。耗时较长，时间复杂度较高并且只适用于二元类不平衡问题 优点：使用决策树基类分类器的 MOSOB 是最稳定和最准确的模型，在静态情况下，在 G 均值和少数类召回方面优于 OOB，并且比 UOB 更能抵抗类不平衡变化
OOB and UOB	Gearbox、Smart Building、PAKDD、iNemo、KDD	MOSOB，RLSACP,WOS-ELM	缺点：只适用于二元类不平衡问题，没有考虑多元，没有概念漂移检测的机制 优点：提出了新的在线类不平衡框架，并且根据类不平衡状态设计了两种重采样策略，方法对少数群体的准确性和整体表现都有效
AWE	Credit Card Fraud Data	Ensemble Naive Bayesian Ensemble RIPPER, Ensemble Decision Tree	缺点：早期最经典的算法，只能适应潜在的概念漂移和少量数据 优点：提出了分类器集成方法新的解决路径，比单分类器提供了更准确的性能
OAUE	Airlines、PAKDD、Poker、Power、Wave	AWE、ACED、DDM、RF、RBF LED	缺点：没有考虑基分类器的多样性对集成模型的影响 优点：全部对比的算法中，内存消耗是少的，并且提出线性函数在快速漂移流上表现更好，但非线性函数对噪声更强
AUE	ELEC、OZONE、DON	HOT、AWE、AUE、HT+WIN	缺点：需要较长恒定的处理时间和内存。对概念漂移的发生不具备良好的恢复性和适应性，并没有增量的学习算法 优点：只是比 AWE 更准确
AUE2	HYP、RBF、SEA、TREE、LED、ELEC、COV、POKER、Airlines	ACE、AUE1、AWE、HOT、DDM、WIN、LEV、NB、OZA、DWM、NSE	缺点：并没有全部采用增量学习方式 优点：AUE2 可以适用于涉及多种漂移和静态环境的场景。AUE2 提供了最佳的平均分类精度，同时证明比其他整体方法消耗更少的内存
CUE	Forest COV、Waveform	AWE	缺点：CUE 算法的训练时间很长，而且在相同数据集上测试的时间一致 优点：准确率和对概念漂移的适应程度要比 AWE 优
ECISD	垃圾邮件数据集、SEA、ELEC	VFDT、AUE2、Learn++、NSE	缺点：实验对比参数只有 G-mean,虽然在和其他三个算法相比，G-mean 只是大体上呈现上升趋势，需要加入更多的参数来分析算法性能 优点：可处理类不平衡的数据流
DWM	SEA	DWM-ITI、DWM-NB	缺点：概念漂移的恢复时间较长 优点：DWM 保持了相当数量的专家，但实现了更高的预测准确度，并更快地收敛到这些准确度
M3	Benchmark PAMAP2、KDDCup'99、 Forest Cove	LB、AHOT、HAT、DWM TAC(时间增强分类器)	缺点：M3 算法在整体分类精度方面排名第三，M3 算法通常仅比主导方法低几个百分点 优点：随着训练数据量的减少，M3 方法开始占主导地位，这表明当存在有限的训练数据时具有更高的精度
DEA	Rand Tree、Waveform、,LED、KDD、spam、SEA、CHESS、ELEC、COV、Rialto 等	SAMKNN、ADACC、DWM、AUE 等	缺点;算法效率较高 优点：在全部的 25 个数据集上的表现，准确度都比其他算法优秀，精度高，而且，能完全适应概念漂移的变化，提出了一种全新的思路设计（超球体）集成分类器
SEA	RTS、RTC、SEA1、SEA-2、SEA3、SEA4、AGRAWAL-1 等	DWM、ASHT、Bagging、ADWIN、Bagging	缺点：如果没有发生概念漂移且数据集较小时，精度明显不如其他算法 优点：当数据集包含突然或逐渐漂移并且处理时间受到关注的问题时，SEA 时最好的选择
SEA2	RTS、RTC、SEA-1、SEA2、SEA3、SEA-4、AGRAWAL1、COV、ELEC、SPAM、AIRL	ADWIN Bagging、ASHT Bagging、Leveraging bagging 等	缺点：网络周期和训练时间较长，当数据越多时，需要建立的网络越密集 优点：增强了启发式规则，在实验的一些数据集上都达到了最优的精度，且内存消耗是最少的
CRHTUS	Forest COV、Poker-Hand、Electricity	ADWIN Bagging	缺点：只适用于少量属性的数据集，预选属性子集和修剪技术还需要进一步完善 优点：采用列属性进行训练，提出了训练新的思路（列属性），并且采用 ADWIN 检测机制，能有效的检测概念漂移
ACE	recurring context data, which was showed in FLORA	WCEA、SEA	缺点：早起的学习算法并不能有效的解决数据块大小和内存消耗的限制，ACE 也是如此，并且没有实现修剪方法 优点：更能抵抗噪声和快速恢复各种类型的概念漂移
WEOB	Gaussian、Gearbox、Smart Building	RLSACP、WOS-ELM	缺点：缺少概念漂移的数据流研究，且都依赖于计算实验数据而得出的结论 优点：改进了 OOB 和 UOB 中的重采样策略，并研究了它们在静态和动态数据流中的性能它们在回忆和 G-均值方面实现了高性能，并且对类不平衡状态的变化表现出良好的稳健性。特别是，WEOB2 显示出比 WEOB1 更好的 G-均值

3.2.2 滑动窗口

滑动窗口在某种意义上类似于地标窗口，它们都定义了窗口大小 n，尽管滑动窗口一次只丢弃一个实例。基于实例的分类器[76~78]。在多集成窗口中 multi-window based ensemble learning(MWEL)[79]定义了三种类型的窗口存储数据流中的最新的实例。窗口类型包括最新的实例窗口和一个集成分类器（包含两个窗口）。集成分类器是由最新的基准分类器和用于训练的每个基准分类器组成。（也就是说，集成分类器首先是由数据流中最新的实例训练得到的一个子分类器和一个由训练数据流中实例的子分类器组成）。这是定义的三种形式的窗口。在预测新到达实例的所属标签以前，对全部的子分类器进行加权操作，当且仅当子分类器的精度低于定义的阈值时才继续训练子分类器。因为如果精度低于或者等于定义的阈值，那么说明当前子分类器的分类性能是和当前数据流中的数据分布背道而驰。

3.2.3 自适应窗口

自适应窗口模型可以被视为具有不同 n 值的标志性窗口。假设流包含具有不同程度和速率的漂移，使用不同大小的窗口是合适的策略。问题是如何根据流的观察动态调整 n。FLORA2[80]算法使用启发式（窗口调整算法）来增加或缩小窗口大小，这是基于另一种猜测漂移是否已经发生的启发式算法。这种用于调整窗口大小的方法在实践中可能是有用的，但是它取决于固定阈值以通过“应该减少”或“增加大小”来定义。最重要的是，它依赖于启发式来确定当前概念是稳定还是发生漂移。ADWIN Bagging 和 Leveraging Bagging 。两种算法都使用 ADWIN（Adaptive Window）漂移探测器来选择性地重置分类器。具体地，在这些算法中，每当其关联的 ADWIN 检测器发出漂移已经发生时，就重置分类器。因此，整体可能最终得到具有对当前概念的不同水平的分类器。ADWIN 算法的主要思路是：最新窗口 W 的两个子窗口 w_1， w_2 可以显示出明显的平均数，并且推断出对应的预测值是相异的，则删除旧窗口。根据 Hoeffding 边界定义两个窗口的平均值大于阈值 ε_{cut}，如公式所示。其中|w|是最新窗口的尺寸，$|w_1|$和$|w_2|$是最新窗口的两个子窗口的尺寸，并且$|w|=|w_1|+|w_2|$。m 是两个子窗口的调和平均数。

$$m=\frac{2}{\frac{1}{|W1|}+\frac{1}{|W2|}} \qquad \varepsilon_{cut}=\sqrt{\frac{1}{2m}\ln\frac{4|W|}{\delta}}$$

3.2.4 地标窗口

地标窗户使用标记方式将数据流分离为互不相交的不同数据块。每当新的实例到达地标时，之前数据块的全部实例将全部被抛弃。通常集成分类器使用固定大小为 n 的标志性窗口来控制集成模型更新周期性。例如分类器的删除，重置，添加或统计重置。这种方法首先在流式集成算法 SEA 中引入，后来用于其他算法，如 dynamic weighted majority（DWM）、accuracy update ensemble（AUE）、accuracy update ensemble2（AUE2）、social adaptive ensemble（SEA）、advances on the social adaptive ensemble（SAE2）、online accuracy update ensemble（OAUE）等。许多用于数据流的集成分类算法都组合了标志性窗口和增量的基础学习器（如 Hoeffding tree）。这种设计选择可以允许合理快速地适应突然漂移（给定小的 n 值），同时它允许集成成员的增量更新。固定的地标窗口方法允许使用传统的批量学习算法进行流学习。在这种情况下，批处理学习器在窗口 w 的实例上进行训练，其模型用于对下一个窗口 $w+1$ 中的实例进行分类。在窗口 $w+1$ 结束后，在 w 上学习的模型被训练的模型 $w+1$ 替换。如果这种方法用于使批量学习器适应流学习，那么可能会出现一些问题，最显着的是：训练集中在窗口之间的过渡期，因此如果新实例快速到达，则有必要考虑预测在训练新模型时出现延误;批量学习者通常需要对大量数据进行训练以获得准确的模型，因此窗口必须非常大，否则学习模型会很弱。最后，如果发生概念漂移，则在窗口结束并生成新模型之前不会考虑它，因此适应突然漂移将是缓慢的。尽管使用固定地标窗口的简单性，但难以定义地标尺寸参数 n。

4 进一步的研究方向

虽然目前研究人员提出了许多数据流集成分类算法，可以解决大部分分类问题，但是还有很多目前不能解决的问题，例如新颖类别检测，多类标检测等问题。况且在带有突变和重现概念漂移的情况下，如何能提高分类器的性能，如何让分类器具有快速的适应能力和恢复能力都是本文以后研究的主要方向。

a)首先在新颖类别检测上，针对可探测新颖类别的数据流集成分类算法不能处理混合属性且新颖类别探测准确率不高的问题，拟采用 AUE2 作为基准算法，并改进新颖类别探测方法以处理混合属性数据和提高新颖类别的探测准确率。根据这样的假设，相同类标实例的距离比其他类标的实例要远，利用空间所占的属性比例来判断新颖类标的存在范围，因为如果是新颖类标的话，那么它必然会落入到另一个区域，同时满足高内聚性的特点。

b)在多类标检测问题中，集成方法显然表现出了比单分类器模型更好的性能，因此本文准备用集成的方法来解决这个问题。考虑都否把多个类别标签放在一个集合中，采用集成模型中权重的方式，对集合中的类别标签进行加权，在预测未知样本的类标时，集成模型给出一个最可能的类标，最后把多类标问题转换成了单一类标。但是这么做，会考虑如何更新在集合中全部的类标，目前准备还是采用内聚性的特点来解决这个问题，还需要进一步的研究和实验。

c)在重现概念漂移中最难解决的问题就是，如何判断新到来的概念是否是学习器之前学习过的概念，因此将致力于解决这一块问题。考虑能否使用多集成窗口模型来解决这个问题。基于这样的实验研究思路：在子集成窗口中，当最新数据流中的概念流入到窗口中，判断是否具有新颖类标，具有新颖类标那么一定是之前没有学习过的，当窗口中长期没有发生的概念



突然重现了，那么在多集成窗口中添加漂移检测和重现漂移检测，新颖类标检测，遗忘机制的做法来解决这个问题。

5　结束语

本文对现有的 40 多种数据流集成分类算法进行综述，详细介绍了各种算法和适用的各种环境。分析了算法的优缺点，实验数据集和对比算法。在最后介绍了进一步研究的方向和需要解决的问题，提出了研究思路和问题解决的办法。

参考文献：

[1] Aloraini A. Penalized ensemble feature selection methods for hidden associations in time series environments case study: equities companies in Saudi Stock Exchange Market [J]. Evolving Systems, 2014, 6 (2): 1-8.

[2] Alzoubi O, Fossati D, D'Mello S, *et al.* Affect detection from non-stationary physiological data using ensemble classifiers [J]. Evolving Systems, 2015, 6 (2): 79-92.

[3] Ditzler G, Roveri M, Alippi C, *et al.* Learning in Nonstationary Environments: A Survey [J]. IEEE Computational Intelligence Magazine, 2015, 10 (4): 12-25.

[4] Amiribesheli M, Benmansour A, Bouchachia A. A review of smart homes in healthcare [J]. Journal of Ambient Intelligence & Humanized Computing, 2015, 6 (4): 495-517.

[5] Bifet A, Read J, Pfahringer B, *et al.* Pitfalls in benchmarking data stream classification and how to avoid them [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. New York: Springer-Verlag Inc. , 2013: 465-479.

[6] Lu Zhenyu, Wu Xindong, Bongard J C. Active learning through adaptive heterogeneous ensembling [J]. IEEE Trans on Knowledge & Data Engineering, 2014, 27 (2): 368-381.

[7] Haque A, Parker B, Khan L. Labeling Instances in evolving data streams with MapReduce [C]// Proc of IEEE International Congress on Big Data. 2013: 387-394.

[8] Khamassi I, Sayed-Mouchaweh M, Hammami M, *et al.* Discussion and review on evolving data streams and concept drift adapting [J]. Evolving Systems, 2016, 9 (1): 1-23.

[9] Vivekanandan P, Nedunchezhian R. Mining data streams with concept drifts using genetic algorithm [J]. Artificial Intelligence Review, 2011, 36 (3): 163-178.

[10] 毛国君，胡殿军，谢松燕. 基于分布式数据流的大数据分类模型和算法 [J]. 计算机学报，2017 (1): 161-175. (Mao Guojun, Hu Dianjun, Xie Songyan. Big data classification model and algorithm based on distributed data stream [J]. Chinese Journal of Computers, 2017 (1): 161-175)

[11] Chapelle O, Chapelle O, Langford J. A reliable effective terascale linear learning system [J]. Journal of Machine Learning Research, 2014, 15 (1): 1111-1133.

[12] Yousefi M, Yousefi M, Ferreira R, *et al.* Chaotic genetic algorithm and Adaboost ensemble metamodeling approach for optimum resource planning in emergency departments. [J]. Artificial Intelligence in Medicine, 2017, 84: 23-33.

[13] Barddal J P, Gomes H M, Enembreck F. A survey on feature drift adaptation [C]// Proc of IEEE International Conference on Tools with Artificial Intelligence. Washington DC: IEEE Computer Society, 2015: 1053-1060.

[14] Biggio B, Corona I, Nelson B, *et al.* Security evaluation of support vector machines in adversarial environments [M]// Support Vector Machines Applications. Springer, 2014: 105-153.

[15] Lu Yanyun, Boukharouba K, Boonært J, *et al.* Application of an incremental SVM algorithm for on-line human recognition from video surveillance using texture and color features [J]. Neurocomputing, 2014, 126 (3): 132-140.

[16] Saffari A, Leistner C, Santner J, *et al.* On-line Random Forests [C]// Proc of IEEE, International Conference on Computer Vision Workshops. 2009: 1393-1400.

[17] Losing V, Hammer B, Wersing H. Interactive online learning for obstacle classification on a mobile robot [C]// Proc of International Joint Conference on Neural Networks. 2015: 1-8.

[18] Barddal J P, Gomes H M. SFNClassifier: a scale-free social network method to handle concept drift [S]. 2014: 786-791.

[19] 周志华. 机器学习 Machine learning [M]. 北京：清华大学出版社, 2016. (Zhou Zhihua. Machine Learning Machine learning [M]. Beijing: Tsinghua University Press, 2016.)

[20] Abbaszadeh O, Amiri A, Khanteymoori A R. An ensemble method for data stream classification in the presence of concept drift [J]. Frontiers of Information Technology & Electronic Engineering, 2015, 16 (12): 1059-1068.

[21] 丁剑，韩萌，李娟. 概念漂移数据流挖掘算法综述 [J]. 计算机科学, 2016, 43 (12): 24-29. (Ding Jian, Han Meng, Li Juan. Overview of conceptual drift data stream mining algorithms [J]. Computer Science, 2016, 43 (12): 24-29.)

[22] Krawczyk B, Cano A. Online ensemble learning with abstaining classifiers for drifting and noisy data streams [J]. Applied Soft Computing, 2018, 68: 677-692

[23] 文益民，强保华，范志刚. 概念漂移数据流分类研究综述 [J]. 智能系统学报, 2013, 8 (2): 95-104. (Wen Yimin, Qiang Baohua, Fan Zhigang. A review of research on conceptual drift data stream classification [J]. Journal of Intelligent Systems, 2013, 8 (2): 95-104.)

[24] Czarnowski I, Jędrzejowicz P. Ensemble classifier for mining data streams [J]. Procedia Computer Science, 2014, 35 (9): 397-406.

[25] Bosnić Z, Demšar J, Kešpret G, *et al.* Enhancing data stream predictions with reliability estimators and explanation [J]. Engineering Applications of Artificial Intelligence, 2014, 34 (3): 178-192.

[26] Wozniak M. A hybrid decision tree training method using data streams [M]. New York: Springer-Verlag Inc. 2011.

[27] Shan Jingsong, Luo Jianxin, Ni Guiqiang, *et al.* CVS: Fast cardinality



estimation for large-scale data streams over sliding windows [J]. Neurocomputing, 2016, 194 (1): 107-116.

[28] Bifet A, Holmes G, Pfahringer B, *et al.* New ensemble methods for evolving data streams [C]// Proc of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2009: 139-148.

[29] Bifet A, Holmes G, Pfahringer B. Leveraging Bagging for Evolving Data Streams [C]// Proc of European Conference on Machine Learning and Knowledge Discovery in Databases. Springer-Verlag, 2010: 135-150.

[30] Sakthithasan S, Pears R, Koh Y S. One pass concept change detection for data streams [C]// Proc of Pacific-Asia Conference on Knowledge Discovery and Data Mining. Berlin: Springer, 2013: 461-472.

[31] Pears R, Sakthithasan S, Koh Y S. Detecting concept change in dynamic data streams [J]. Machine Learning, 2014, 97 (3): 259-293.

[32] 白洋. 数据流概念漂移检测和不平衡数据流分类算法研究 [D]. 北京: 北京交通大学, 2017. (Bai Yang. Research on data stream concept drift detection and unbalanced data stream classification algorithm [D]. Beijing: Beijing Jiaotong University, 2017.)

[33] 李南, 郭躬德, 陈黎飞. 基于少量类标签的概念漂移检测算法 [J]. 计算机应用, 2012, 32 (8): 2176-2181. (Li Nan, Guo Yude, Chen Lifei. Concept drift detection algorithm based on a small number of class labels [J]. Journal of Computer Applications, 2012, 32 (8): 2176-2181.)

[34] Mejri D, Khanchel R, Limam M. An ensemble method for concept drift in nonstationary environment [J]. Journal of Statistical Computation & Simulation, 2013, 83 (6): 1115-1128.

[35] Trawiński B, Smętek M, Lasota T, *et al.* Evaluation of Fuzzy System Ensemble Approach to Predict from a Data Stream [M]// Intelligent Information and Database Systems. Springer International Publishing, 2014: 137-146.

[36] Khamassi I, Sayed-Mouchaweh M. Drift detection and monitoring in non-stationary environments [C]// Evolving and Adaptive Intelligent Systems. IEEE, 2014: 1-6.

[37] Barddal J P, Gomes H M, Enembreck F. A survey on feature drift adaptation [C]// Proc of IEEE International Conference on TOOLS with Artificial Intelligence. Washington DC: IEEE Computer Society, 2015: 1053-1060.

[38] Han Donghong, *et al.* Two birds with one stone: classifying positive and unlabeled examples on uncertain data streams [J]. Neurocomputing 277 (2018): 149-160.

[39] 孙艳歌, 王志海, 原继东, 等. 数据流滑动窗口方式下的自适应集成分类算法 [J]. 北京交通大学学报, 2016, 40 (5): 9-15. (Sun Yange, Wang Zhihai, Yuan Jidong, *et al.* Adaptive integrated classification algorithm for data stream sliding window mode [J]. Journal of Beijing Jiaotong University, 2016, 40 (5): 9-15.)

[40] Ikonomovska E, Gama J, Džeroski S. Learning model trees from evolving data streams [J]. Data Mining & Knowledge Discovery, 2011, 23 (1): 128-168.

[41] Bifet A. SAMOA: scalable advanced massive online analysis [M]. JMLR. org, 2015.

[42] Ditzler G, Roveri M, Alippi C, *et al.* Learning in Nonstationary Environments: A Survey [J]. IEEE Computational Intelligence Magazine, 2015, 10 (4): 12-25.

[43] 姜爱克, 赵峰, 张杰. 基于自适应集成分类器的数据流概念漂移算法 [J]. 统计与决策, 2016 (7): 13-17. (Jiang Aike, Zhao Feng, Zhang Jie. Data flow concept drift algorithm based on adaptive integrated classifier [J]. Statistics and Decision, 2016 (7): 13-17.)

[44] Sidhu P, Bhatia M P S. A novel online ensemble approach to handle concept drifting data streams: diversified dynamic weighted majority [J]. International Journal of Machine Learning &Cybernetics, 2015 (1): 1-25.

[45] Wang Haixun, Han Jiawei, *et al.* Mining Concept-Drifting Data Streams [M]// Data Mining and Knowledge Discovery Handbook. 2009: 789.

[46] Street W N, Kim Y S. A streaming ensemble algorithm (SEA) for large-scale classification [C]// Proc. of ACM SIGKDD Int. Conference on Knowledge Discovery and Data Mining. 2001: 377-382.

[47] Ramamurthy S, Bhatnagar R. Tracking recurrent concept drift in streaming data using ensemble classifiers [C]// Proc of International Conference on Machine Learning and Applications. Washington DC: IEEE Computer Society, 2007: 404-409.

[48] Pelossof R, Jones M, Vovsha I, *et al.* Online coordinate boosting [C]// Proc of IEEE, International Conference on Computer Vision Workshops. 2008: 1354-1361.

[49] Brzezinski D, Stefanowski J. Combining block-based and online methods in learning ensembles from concept drifting data streams [J]. Information Sciences, 2014, 265 (5): 50-67.

[50] Ludmila I. Kuncheva. A Bound on Kappa-Error Diagrams for Analysis of Classifier Ensembles [J]. IEEE Trans on Knowledge and Data Engineering, 2013, 25 (3): 494-501.

[51] Ikonomovska E, Gama J, Džeroski S. Online tree-based ensembles and option trees for regression on evolving data streams [J]. Neurocomputing, 2015, 150: 458-470.

[52] Wang Boyu, Joelle Pineau. Online bagging and boosting for imbalanced data streams [J]. IEEE Trans on Knowledge & Data Engineering 1 (2016): 1.

[53] Bühlmann P. Bagging, boosting and ensemble methods [M]// Handbook of Computational Statistics. Berlin: Springer, 2012: 985-1022.

[54] Bifet A. Adaptive learning from evolving data streams [C]// Proc of International Symposium on Intelligent Data Analysis: Advances in Intelligent Data Analysis Viii. Springer-Verlag, 2009: 249-260

[55] Žliobaitė I, Bifet A, Read J, *et al.* Evaluation methods and decision theory for classification of streaming data with temporal dependence [J]. Machine Learning, 2015, 98 (3): 455-482.

[56] 马宪哲. 基于集成分类器的数据流分类算法研究 [D]. 长春: 东北大学, 2012. (Ma Xianzhe. Research on data stream classification algorithm based on ensemble classifier [D]. Changchun: Northeastern University, 2012.)



[57] Kourtellis N, DeMorales G F, Bifet A, *et al.* VHT: vertical hoeffding tree [C]// Proc of IEEE International Conference on Big Data. Piscataway, NJ: IEEE Press, 2016: 915-922. 2016: 915-922.

[58] Wang Shuo, Minku, Leandro L, Yao Xin. A multi-objective ensemble method for online class imbalance learning. [C]// Proc of International Joint Conference on Neural Networks. 2014. p. 3311-3318.

[59] Wang Shuo, Minku, Leandro L, Yao Xin. A learning framework for online class imbalance learning [C]// Proc of IEEE Symposium on Computational Intelligence and Ensemble Learning. 2013: 36-45.

[60] Littlestone N, Warmuth M K. The weighted majority algorithm (supersedes 89-16) [J]. Revista Española De Física, 2011.

[61] Brzezinski D, Stefanowski J. Reacting to different types of concept drift: the Accuracy Updated Ensemble algorithm [J]. IEEE Trans on Neural Networks & Learning Systems, 2014, 25 (1): 81-94.

[62] Md Farid D, Zhang Li, Hassain A, *et al.* An adaptive ensemble classifier for mining concept drifting data streams [J]. Expert Systems with Applications, 2013, 40 (15): 5895-5906.

[63] Wang Shuo, Minku L L, Yao Xin. Resampling-based ensemble methods for online class imbalance learning [J]. IEEE Trans on Knowledge and Data Engineering, 2015, 27 (5): 1356-1368.

[64] Krawczyk B, Cano A. Online ensemble learning with abstaining classifiers for drifting and noisy data streams [J]. Applied Soft Computing, 2018, 68: 677-692.

[65] Parker B S, Khan L, Bifet A. Incremental ensemble classifier addressing non-stationary fast data streams [C]// Proc of IEEE International Conference on Data Mining. 2014: 716-723.

[66] Loeffel P X, Bifet A, Marsala C, *et al.* Droplet Ensemble Learning on Drifting Data Streams [C]// Proc of International Symposium on Intelligent Data Analysis. Cham: Springer, 2017: 210-222

[67] Abbaszadeh O, Amiri A, Khanteymoori A R. An ensemble method for data stream classification in the presence of concept drift [J]. Frontiers of Information Technology & Electronic Engineering, 2015, 16 (12): 1059-1068.

[68] Bifet A, Frank E, Holmes G, *et al.* Accurate ensembles for data streams: combining restricted Hoeffding trees using stacking [J]. Journal of Machine Learning Research, 2010, 13 (13): 225-240.

[69] Gomes H M, Enembreck F. SAE: social adaptive ensemble classifier for data streams [C]// Computational Intelligence and Data Mining. 2013: 199-206.

[70] Gomes H M, Enembreck F. SAE2: advances on the social adaptive ensemble classifier for data streams [C]// Proc of ACM Symposium on Applied Computing. 2014.

[71] Wozniak M, Ksieniewicz P, Cyganek B, *et al.* Active learning classification of drifted streaming data [J]. Procedia Computer Science, 2016, 80 (C): 1724-1733.

[72] Yin Chunyong, Feng Lu, Ma Luyu. An improved Hoeffding-ID data-stream classification algorithm [J]. Journal of Supercomputing, 2016, 72 (7): 2670-2681.

[73] Sarafis I, Diou C, Delopoulos A. Online training of concept detectors for image retrieval using streaming clickthrough data [J]. Engineering Applications of Artificial Intelligence, 2016, 51: 150-162.

[74] Sancho-Asensio A, Orriols-Puig A, Casillas J. Evolving association streams [J]. Information Sciences, 2016, 334–335 (C): 250-272.

[75] Ryang H, Yun U. High utility pattern mining over data streams with sliding window technique [J]. Expert Systems with Applications, 2016, 57: 214-231.

[76] Naik S B, Pawar J D. A quick algorithm for incremental mining closed frequent itemsets over data streams [C]// Proc of ACM IKDD Conference on Data Sciences. New York: ACM Press, 2015: 126-127.

[77] Lifna C S, M. Vijayalakshmi D. Identifying concept-drift in Twitter streams [J]. Procedia Computer Science, 2015, 45: 86-94.

[78] Wang Ye, Li Hu, Wang Hua, *et al.* Multi-window based ensemble learning for classification of imbalanced streaming data [C]// Proc of International Conference on Web Information Systems Engineering. 2015: 78-92.

[79] Widmer G, Kubat M. Learning flexible concepts from streams of examples: FLORA2 [C]// Proc of European Conference on Artificial Intelligence. Hoboken: Wiley, 1992: 463-467.